Maturation of naïve and antigen-experienced B-cell receptor repertoires with age

Marie Ghraichy[1,2], Jacob D. Galson[2], Aleksandr Kovaltsuk[3], Valentin von Niederhäusern[1,2], Jana Pachlopnik Schmid[1,2], Enkelejda Miho[4], Dominic F. Kelly[5], Charlotte M. Deane[3], Johannes Trück[1,2*]

[1]Division of Immunology, University Children's Hospital, University of Zurich, Zurich, Switzerland

[2]Children's Research Center, University of Zurich, Zurich, Switzerland

[3]Department of Statistics, University of Oxford, Oxford, United Kingdom

[4]University of Applied Sciences and Arts Northwestern Switzerland FHNW, Institute of Medical Engineering and Medical Informatics, Muttenz, Switzerland

[5]Oxford Vaccine Group, Department of Paediatrics, University of Oxford, Oxford, United Kingdom

[*]Corresponding author: Johannes Trück, MD DPhil, University Children's Hospital, Steinwiesstrasse 75, 8032 Zurich, Switzerland. +41 44 266 7111; johannes.trueck@kispi.uzh.ch

Keywords: Antibody; B cells; B-cell receptor; Children; Maturation; Repertoire

Content:

- Figures: 7
- Tables: -
- Supplementary Figures: 8
- Supplementary Tables: 2
- Word count manuscript: 4445
- References: 59

Funding:
Swiss National Science Foundation (Ambizione-SCORE: PZ00P3_161147; PZ00P3_183777) (JT)
Gottfried und Julia Bangerter-Rhyner-Stiftung (JT)
Olga Mayenfisch Stiftung (JT)
Investment fund of the University of Zurich (JT)

Abstract

B cells play a central role in adaptive immune processes, mainly through the production of antibodies. The maturation of the B-cell system through continuous antigen exposure with age is poorly studied. We extensively investigated naïve and antigen-experienced B-cell receptor (BCR) repertoires in individuals aged 6 months to 50 years. Most dynamics were observed in the first 10 years of life characterized by an increase in frequencies of mutated transcripts through positive selection, increased usage of downstream constant region genes and a decrease in the frequency of transcripts with self-reactive properties. Structural analysis revealed that the frequency of antibodies different from germline in shape increased with age. Our results suggest large and broad changes of BCR repertoires through childhood and stress the importance of using well-selected, age-appropriate controls in BCR studies.

(130 words)

Introduction

B cells play a central role in physiological adaptive immune processes and exert their main effector function through production of antibodies (1). B cells also contribute to the pathogenesis of autoimmune disease via generation of auto-reactive antibodies and modulation of T-cell responses (2, 3). The heavy and light chains of the B-cell receptor (BCR, membrane-bound antibody) are generated in the bone marrow by recombining individual variable (V), diversity (D) and joining (J) gene segments through a process called VDJ recombination. Upon antigen recognition, a BCR is further diversified through rounds of somatic hypermutation (SHM) leading to affinity maturation whereby B cells with improved antigen-binding properties are selected in the germinal center. Class switch recombination (CSR) is also initiated following antigen encounter, causing a change in the constant region of the BCR and a change in its effector function.

Detailed characterization of B cells and their respective BCR sequences offers important information on B-cell generation and selection as well as immune competence in health and disease. High-throughput sequencing of antibody genes (Ig-seq) has become a widely used tool in human translational research (4, 5). Abnormal B-cell responses can be explored by investigating BCR repertoires from patients and comparing their characteristics to those of healthy controls. The limited data already available suggests that it is likely that significant changes occur in the properties of BCR repertoires with age (6). It is therefore important to establish robust data on normal BCR repertoires within sufficiently narrow age-bands to fully understand the process of BCR maturation. This will facilitate the use of BCR repertoire sequencing to understand changes of relevance to childhood disease. Given the high burden of infectious diseases in childhood and the importance of effective immune response to vaccines to prevent infection this is an important group from which to have normative data. There are very few studies that have used Ig-seq to investigate the healthy BCR repertoire, and these studies are limited in the age range of the participants (7–10). In a more detailed study, Ijspeert et al. reported on the antigen-experienced (i.e. IgA and IgG) BCR repertoires of 38 healthy control (HC) samples with their ages ranging from newborn to 74 years (11). The authors found several characteristics of the studied BCR repertoire varying with age and identified pattern that are specific for isotype subclasses. However, their study was limited by the number of samples from children, the low depth of sequencing, and the B-cell subsets analyzed.

We undertook a study to assess in detail the naïve and antigen-experienced BCR repertoires in children and young adults using advanced Ig-seq technology and extensive bioinformatic analysis. This approach allowed us to comprehensively assess factors affecting the healthy BCR repertoire and also provides a robust data set that can serve as a future reference for studying BCR repertoires in children as well as young adults.

Methods

Study participants and cell isolation

Study participants were recruited with informed consent under approval from the Ethics Committee Zurich (KEK-ZH 2015-0555). Peripheral blood samples (5-9 mL) were collected at a single time point from 46 healthy participants aged 6 months to 50 years (Supplementary table 1). Peripheral blood mononuclear cells (PBMC) were isolated by centrifugation of PBS-diluted blood over Ficoll-Paque Plus (Sigma-Aldrich). Either PBMC or B cells magnetically sorted using the AutoMACS Pro cell separator and CD19+ microbeads (both Miltenyi Biotec) were lysed in RLT buffer (Qiagen), snap frozen on dry ice and then stored at -80 °C prior to use. Cells were counted using an optical microscope and a hemocytometer (improved Neubauer). The B-cell number was recorded based on actual counts or estimated using PBMC counts and either B-cell frequencies from flow cytometry performed on the same blood sample or the median percentage of age-dependent reference values (12) if the former was not available.

RNA isolation and library preparation

RNA was extracted from stored samples using the RNeasy Mini Kit (Qiagen). Reverse transcription was performed using SuperScript III/IV (Invitrogen) according to the manufacturer's instructions and constant region primers that included 14 nt unique molecular identifiers (UMI), and partial p7 adaptors. Two reverse transcription reactions were carried out for each sample: one with a mix of IgM and IgD-specific reverse primers and another with a mix of IgA, IgG, and IgE-specific reverse primers. From 6 samples, one mix with all the different C region primers were used in a single reaction. Primer sequences with concentrations are included in Supplementary table 2. BCR heavy chain rearrangements were amplified in a two-round multiplex PCR; the first round using a mix of V family specific forward primers with partial p5 adaptors, and the second round to complete the adaptor sequences. PCR conditions for the first round were 95 °C for 5 min, either 8 cycles (IgD/IgM) or 12 cycles (IgA/E/G) of 98 °C for 20s, 60 °C for 45s and 72 °C for 1 min, and 72 °C for 5 minutes. The PCR conditions for the second round were 95 °C for 5 min, 22 cycles of 98 °C for 20s, 69 °C for 20s and 72 °C for 15 sec, and 72 °C for 5 minutes. PCR amplicons were gel-extracted, purified and quantified using the Illumina qPCR library quantification protocol. Individual libraries were normalized based on concentration and then multiplexed in batches of 12 for sequencing on the Illumina MiSeq platform (2 x 300 bp paired-end chemistry).

Sequence processing and annotation

Samples were demultiplexed via their Illumina indices, and initially processed using the Immcantation toolkit (13, 14). Briefly, raw fastq files were filtered based on a quality score threshold of 20. Paired reads were joined if they had a minimum length of 10 nt, maximum error rate of 0.3 and a significance threshold of 0.0001. Reads with identical UMI (i.e. originating from the same mRNA molecule) were collapsed to a consensus sequence. Reads with identical full-length sequence and identical constant primer but differing UMI were further collapsed resulting in a dataset containing a set of unique sequences per sample and isotype. Sequences were then submitted to IgBlast (15) for VDJ assignment and sequence annotation, and unproductive sequences removed. Constant region sequences were mapped to germline using Stampy (16), and only sequences with a defined constant region were kept for further analysis. The number and type of V gene mutations was calculated using the shazam R package (14). Selection pressure was calculated using BASELINe (17) implemented within shazam.

Sequence clustering

Sequences were independently clustered for each sample to group together those arising from clonally related B cells. The clustering required identical V and J segment use, identical complementary-determining region CDR3 length, and allowing a 1 in 12 amino acid mismatch in the CDR3 as previously determined (7, 8). Lineages were constructed from clusters using the

alakazam R package (18). For calculation of selection pressure of samples, individual sequences within clusters are not independent events, so an effective representative sequence of each clonal group was determined using the default settings of shazam.

From Sequence to Structure

The pipeline to annotate Ig-seq data with structural information consisted of three steps: filtering and numbering amino acid sequences with ANARCI (19), annotation of antibody canonical loop classes (CDR-H1 and CDR-H2) with SCALOP (20) and mapping the CDR-H3 loop sequences to known antibody structures using FREAD (21, 22).

Antibody sequence filtering and numbering were performed using the ANARCI (19) parsing step as used in the first steps of the ABOSS algorithm (23). Herein, sequences were checked for structural viability. Sequences were filtered out that 1) contained indels in framework regions (FWR) or canonical CDR, 2) could not be aligned to the human Hidden Markov Model profile of an IMGT germline, 3) had a J gene sequence identity less than 50% to a human IMGT germline, and 4) contained non-amino acid entries. Since the primer masking step in pRESTO (13) removed FWR1 and positions 127 and 128 in some sequences, ANARCI parsing was customized to account for these exceptions. Sequences that passed these criteria were numbered using ANARCI (19) with the IMGT scheme (24).

To annotate the numbered sequences with canonical loop classes information, we employed SCALOP (20) using the IMGT CDR definition (24). In SCALOP, CDR-H1 and CDR-H2 canonical loop classes have backbone RMSD thresholds of 0.8Å and 0.63 Å (20, 25). The expected coverage of canonical loop class sequences with SCALOP is 93%, where 89% of predicted templates have RMSD values within 1.5Å of the correct structure. FREAD (21, 22) is used to predict CDR-H3 templates and operates on a pre-built database of antibody structures, returning the optimal structure protein data bank (PDB) code and chain identifier for the queried sequence. The FREAD database was downloaded from SAbDab (26), and consists of all X-ray crystal structures with resolution better than 2.9Å deposited in the PDB before November 14, 2018. As the original SAAB pipeline works with Chothia numbering (27), the ESS length cutoffs were incremented by 2 to map to the IMGT numbering scheme. The expected coverage of redundant CDR-H3 sequences with FREAD is 75%, where 80% of predicted templates have RMSD values within the 1.5Å range to a target structure.

Statistical analysis and graphing

Statistical analysis and plotting were performed using R (28); all plots were produced using the ggplot2 and ggpubr packages (29, 30). Specific tests used are detailed in the figure legends.

Classification of sequences into naive and antigen-experienced subsets using isotype and mutational rate

Using constant region annotation and mutation number, individual sequences were grouped into biologically different subsets based on known B-cell subpopulations. Based on the frequency distribution of mutations for IgD and IgM sequences, those with up to 3 nt mutations across the entire V gene were considered “unmutated” (naïve) to account for allelic variance (31) and remaining PCR and sequencing bias (Supplementary figure 1). All class-switched sequences were defined as antigen-experienced irrespective of their V gene mutation count. Because of very low sequence numbers, IgE and IgG4 transcripts were excluded from most analysis. The number of sequences of the different subsets among total transcripts by individual are found in Supplementary table 1.

Data availability

Raw sequence data used for analysis in this study are available at the NCBI Sequencing Read Archive (www.ncbi.nlm.nih.gov/sra) under BioProject number PRJNA527941 including metadata meeting the MiAIRR standards (32). The processed and annotated final dataset is available in

Zenodo (https://doi.org/10.5281/zenodo.2640393) along with the protocol describing the exact processing steps with the software tools and version numbers.

Results

We obtained 66'920'657 raw sequences from samples of the 46 healthy study participants. Processing, filtering and collapsing of these reads resulted in a total dataset of 7'044'649 unique BCR sequences that were used for downstream analysis. The number of unique sequences per sample correlated and was representative of the number of B cells in that sample (Supplementary figure 2)

V family and J gene usage

Although previous work has observed common patterns of gene segment usage and has suggested a strong dependence on an individual's germline genetic background (33, 34), the relative contributions to variance from age remained unclear. Proportions of sequences assigned to the different V gene families and J genes were calculated for each sample and cell subset. The overall distribution of V family and J gene usage were different in older individuals compared with younger age groups. In particular, frequencies of V1 family sequences significantly decreased with age in naïve and mutated IgD and IgM sequences. This decrease was also observed in IgG and IgA cells with higher individual variation in older subjects (Figure 1A). No clear pattern with statistical significance was found in the usage of the other V families by age (Supplementary figure 3A). Such changes in V1 family genes are due to differences in several different individual V genes, particularly VH1-8 (Supplementary figure 5).

There were also changes in the overall J gene usage over the first 10 years of life marked by a significant decrease in the frequencies of sequences assigned to J6 in IgG subsets (Figure 1B). In naïve, memory IgM/IgD and IgA subsets, a similar trend was observed but was not statistically significant. The frequencies of the remaining J genes by age group are shown in Supplementary figure 3B. According to previous work and to our data (Supplementary figure 3C), BCR sequences with rearranged J6 gene have longer junctions (35, 36). Consistent with this and the decrease of J6 usage with age, a significant decrease in junction length with age was observed in IgG subsets (Figure 1C). However, when normalizing for J gene in IgG subsets, the decrease in junction length with age was still apparent, particularly in J6 transcripts suggesting a decrease of longer J6 usage with age in IgG (Supplementary figure 4)

Somatic hypermutation

Levels of somatic hypermutation (SHM) were determined by calculating V gene mutations in individual sequences, and mean values were calculated across samples and cell subsets. There was a substantial increase in SHM in all antigen-experienced subsets with age, which was most prominent in the first 10 years of life (Figure 2A). Substantial changes in mutation counts were found in all IgA and IgG subsets with exponential increases in children under 10 years and more linear progression between 10 and 50 years. IgD and IgM memory subsets showed the smallest change of all subsets with some increase in children and a plateau from the 2nd decade. However, the proportion of mutated IgM transcripts per sample increased from 0.1 in 0-3 year olds to an average of 0.4 in older individuals. This trend was also observed in IgD transcripts, although statistical significance was not reached (Figure 2B). These changes in IgD and IgM sequences mimicked the age-related trend seen in the proportions of mutated IgA and IgG antigen-experienced sequences, but at a lower level (Figure 2B).

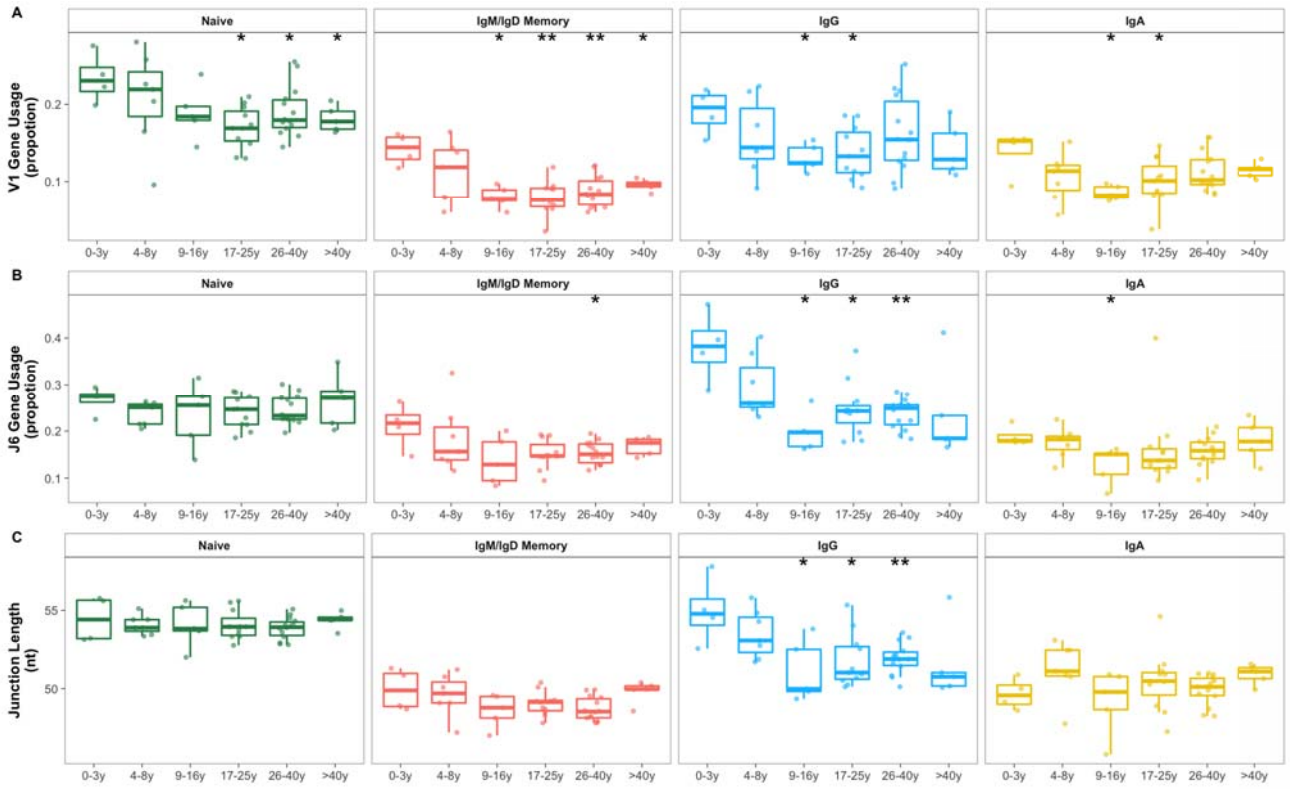


Figure 1: V family and J gene usage is changing in early childhood. *A* V1 family usage is significantly reduced in older individuals compared with young age groups in the different B cell subsets. *B* J6 gene usage significantly decreased in the first 10 years of life in IgG subsets. *C* Mean junction length significantly decreased in the first 10 years of life exclusively in IgG subsets. Comparison of each age group to the 0-3y group was performed using the Wilcoxon test. *p<0.05, **p<0.01

Structural properties

Crystallographic studies have shown that antibody CDR-H1 and CDR-H2 loops can adopt a very limited number of structural conformations, known as canonical loop classes (27, 37). These canonical classes are considered to be separate and distinct structures of the CDR. SCALOP annotates CDR sequences with their canonical loop class rapidly and accurately (~90%) (20). It was used to predict the canonical classes of CDR-H1 and CDR-H2 of the BCR sequences. The proportion of sequences in which either CDR-H1 and CDR-H2 had switched from the canonical class of their germline was calculated and summarized across samples and subsets. In all memory subsets, the proportion of transcripts with structures differing from germline followed a trend similar to the increasing mutation number with age (Figure 2C). In all IgA and IgG subsets, the proportion of sequences different from germline increased exponentially in the first 10 years of life with a more linear increase between 10 and 50 years. This trend was still present but less pronounced in IgD and IgM memory subsets.

Structures of CDR3 were predicted by mapping sequences to a pre-built database of antibody structures. Sequences were annotated with a PDB code identifier. The proportion of every PDB structure within individual and repertoire was calculated, and PDB codes identified that positively or negatively correlated with age or showed no age dependency. Figure 3 shows predicted PDB structures that increased (left) or decreased (right) with age in IgG and IgA repertoires. In naïve and IgM/IgD memory repertoires, similar findings were observed (Supplementary figure 6).

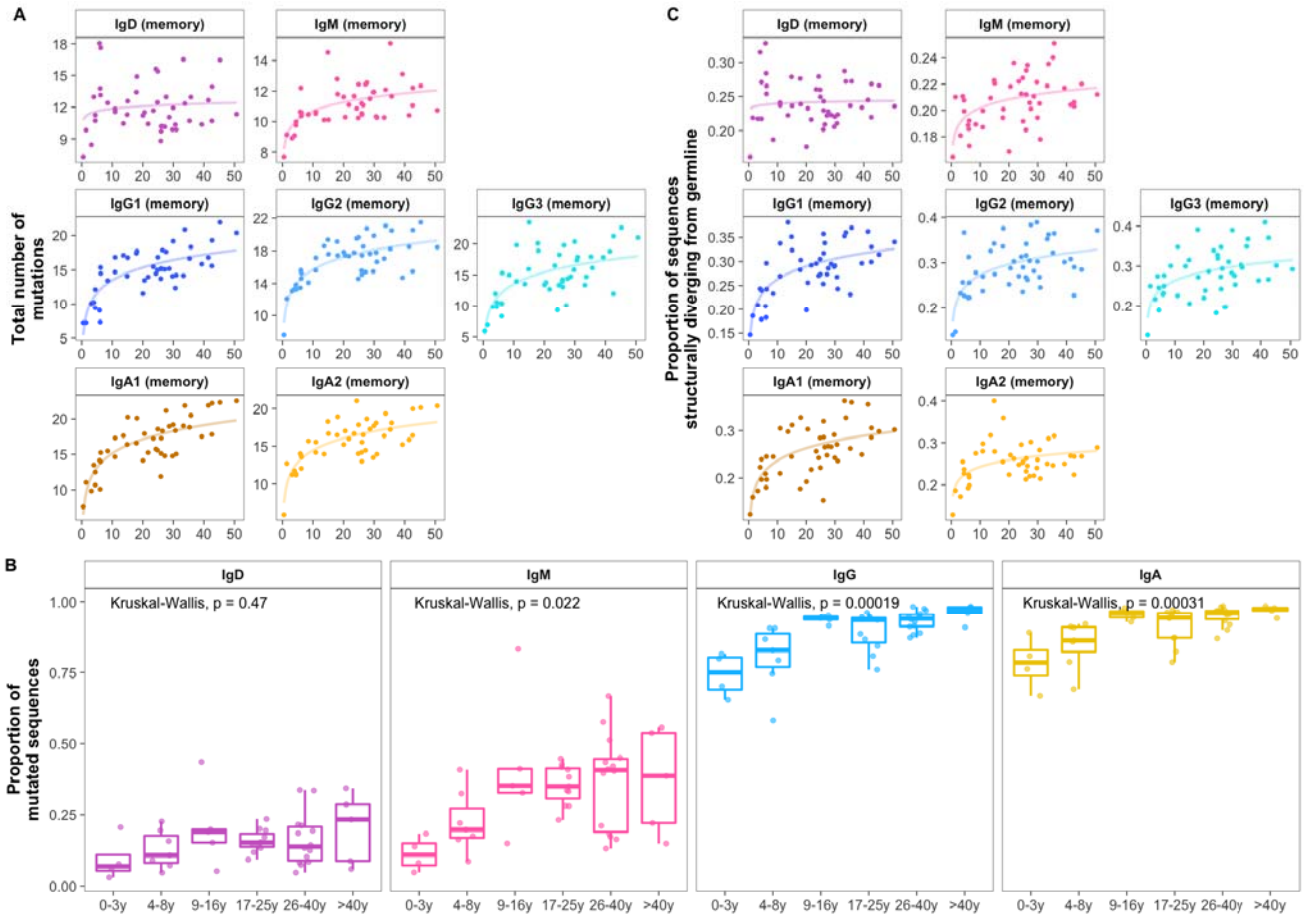


Figure 2: Age-related changes in somatic hypermutation and structure. *A* Mean number of V gene mutations by individual and B-cell subset with fitted logarithmic curves. Somatic hypermutation increased mainly in the first 10 years of life with some differences between cell subsets. *B* The proportion of memory IgD and memory IgM out of all IgD/IgM transcripts and the proportion of mutated IgG and IgA transcripts within repertoires showed significant increases in the first 10 years of life. *C* The proportion of sequences structurally different from germline increased in early childhood in all B-cell subsets. Statistical differences between groups were tested using the Kruskal Wallis test.

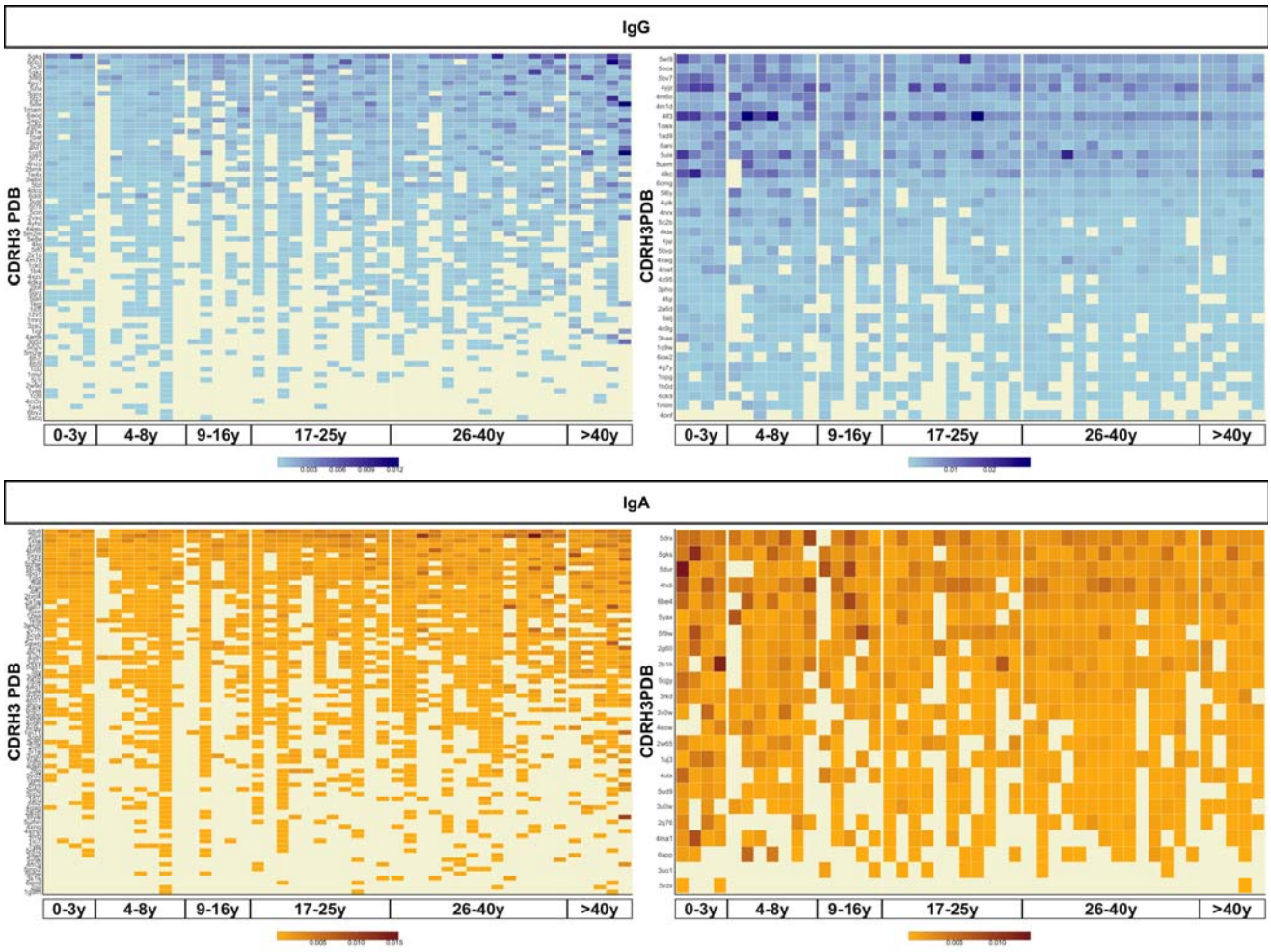


Figure 3 Structural composition of the IgG and IgA repertoire is partly age dependent. PDB codes that have a positive correlation (left) and a negative correlation (right) with age and with a Pearson correlation p-value <0.05 are shown. Out of 1935 unique IgG PDBs, 68 (0.04%) and 38 (0.02%) were positively and negatively correlated with age, respectively. Out of 1907 unique IgA PDBs, 78 (0.04%) and 23 (0.01%) were positively and negatively correlated with age, respectively. Samples are ordered by age and PDB codes are ordered by sharedness across individuals.

Clonal expansion

Lineage trees were constructed from the clusters of clonally expanding cells and used to determine the evolutionary relationship of B cells within the cluster. For each lineage, the trunk length (distance between the most recent common ancestor and germline sequence) was calculated (Figure 4A) – as a measure of the maturity of a lineage (38). There was a significant positive correlation between trunk length and age (Figure 4B). The gini index of each lineage was also calculated, which gives insight into whether the lineage is dominated by a single clone (high gini index) or has a broad branching structure (low gini index). This characteristic showed a negative correlation with age (Figure 4C).

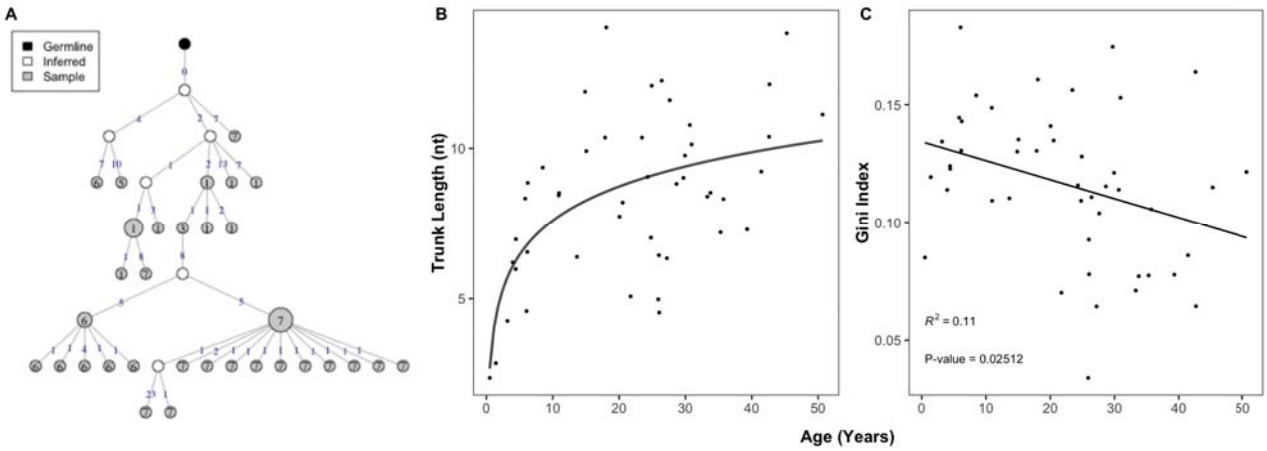


Figure 4: Age-related changes in clonal expansions. *A* Example structure of a lineage tree. Each node is a sequence, and the size of the node represents the number of identical sequences. The numbers on the connectors show the number of mutations that separate the sequences. *B* The correlation between age and mean trunk length with a fitted logarithmic curve. *C* The correlation between mean gini index and age with a fitted linear curve. R-squared and p values of the linear model are shown.

Antigen-driven selection

Insights into the process of antigen-driven selection can be gained by analyzing the mutational pattern in antigen-experienced repertoires. We aimed to determine antigen-driven selection of observed sequences in different ways. First, we calculated the replacement-to-silent (R/S) ratio as a measure of antigen-driven enrichment of sequences with mutations having an effect on the amino acid level vs. those that did not alter the sequence. The R/S ratio was measured separately in FWR and CDR. Sequences with a higher R/S ratio, particularly in CDR, are thought to be under strong antigen selection. Sequences that contained replacement but no silent mutations, the number of silent mutations was set to 1. The R/S ratio in CDR showed a marked increase in all antigen-experienced subsets between 0 and 10 years of life (Figure 5A). In samples from study participants older than 10 years, the R/S ratio was largely constant with values of around 3-3.5 in all B-cell subsets with some variation among individuals. In contrast, the R/S ratio was less variable and lower in FWR compared with CDR and no association with age was found (Supplementary figure 7).

Next, we determined selection pressure using BASELINe – this calculates selection by comparing the observed mutations to expected mutations derived from an underlying SHM targeting model (17). Again, this was conducted separately for CDR and FWR. CDR were under positive selection, whereas FWR were under negative selection. Selection also varied with age, but this relationship was different for the different B-cell subsets. Using this model, there was a decrease in selection pressure for IgA1, IgA2, IgG1, IgG2 and IgG3 with age, and no age-dependent changes for IgD or IgM memory (Figure 5B). The statistical framework used to test for selection was CDR_R / (CDR_R + CDR_S), which normalizes for the observed increase in the total number of mutations with age. The observed decrease in selection pressure in class-switched subsets indicates that young individuals show more dynamics to achieve highly selected sequences compared with older individuals.

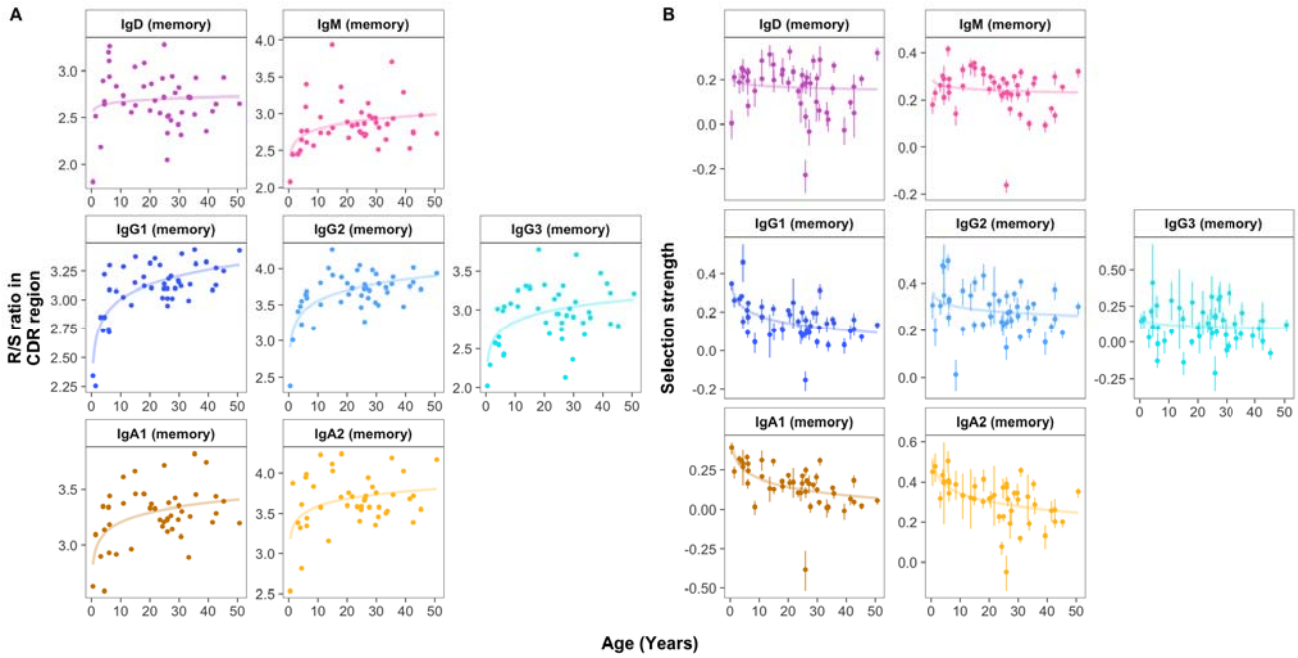


Figure 5: Age-related changes in selection pressure. A) Mean R/S ratio in V gene CDR as a measure of selection pressure showed an increase in early childhood in all B-cell subsets. **B)** Mean selection strength calculated using BASELINE decreases with age in class switched subsets. Error bars represent 95% confidence interval.

Isotype subclass usage

Within IgA and IgG repertoires, the frequency of each isotype subclass was determined, and age-related changes were explored. In all age groups, IgG1 sequences were the most commonly detected, followed by IgG2, IgG3, and IgG4 sequences. However, the proportion of IgG2 sequences increased with age (p=0.021, Kruskal-Wallis) at the expense of lower usage of IgG1 (p=0.012, Kruskal-Wallis) and IgG3 (p=0.16, Kruskal-Wallis) sequences in older people. Similarly, IgA1 was most commonly used in all age groups but the proportion of IgA2 sequences also increased with age (p=0.033, Kruskal-Wallis) (Figure 6).

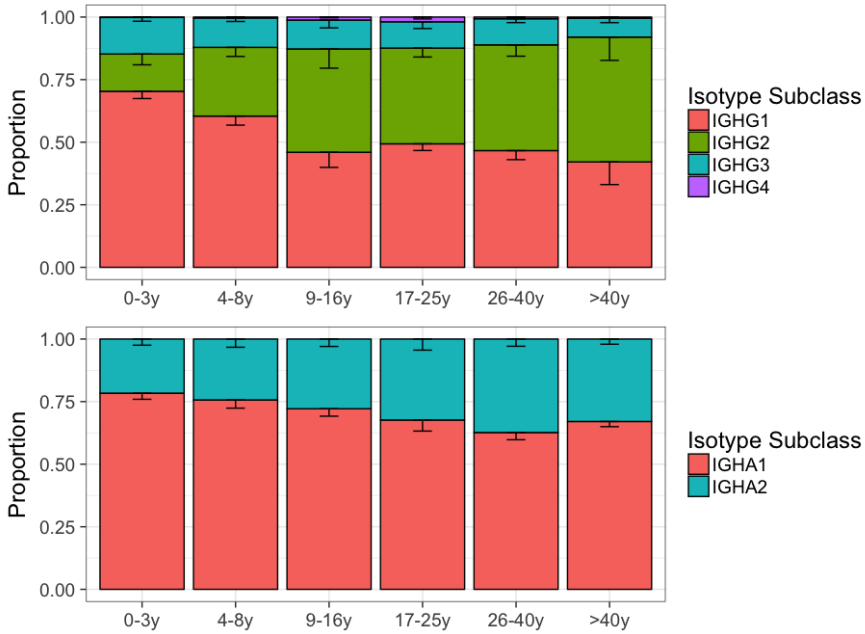


Figure 6: Class switching in healthy controls. The IgG and IgA isotype subclass distribution changes with age. Error bars represent SEM.

Repertoire features associated with autoimmunity

The autoimmune potential of BCR repertoires can be investigated by searching for sequences that demonstrate certain characteristics associated with self-reactivity. These include an increased usage of certain V genes, mainly VH4-34, and usage of longer CDR3 with positively charged or hydrophobic residues (39–41). We investigated how these metrics vary with age in healthy individuals. Apart from the decreasing junction length in IgG subsets (Figure 1C), we found that age has no impact on charge or hydrophobicity (Supplementary figure 8). Overall VH4-34 usage was also unrelated to age whereas a more detailed SHM analysis including self-reactive motifs of VH4-34 sequences revealed an age-specific pattern. The VH4-34 germline contains an Ala-Val-Tyr (AVY) hydrophobic patch in the FWR1 that is not present in other V segments and is thought to contribute to the self-reactive property of this gene (42, 43). Another feature of the VH4-34 germline associated with autoimmunity is the presence of an Asn-X-Ser N-glycosylation sequon (NHS) in the CDR2 that modulates antibody avidity (44). Previous research has shown that mutating one or both of these motifs drives specificity of these sequences away from self, thereby contributing to peripheral tolerance. Lower frequencies of both unmutated AVY and NHS were present in healthy cohorts of older people while there was a relative accumulation of single and double-mutated motifs in VH4-34 with age (Figure 7). This pattern was observed across all antigen-experienced subsets but was more pronounced in IgA and IgG transcripts.

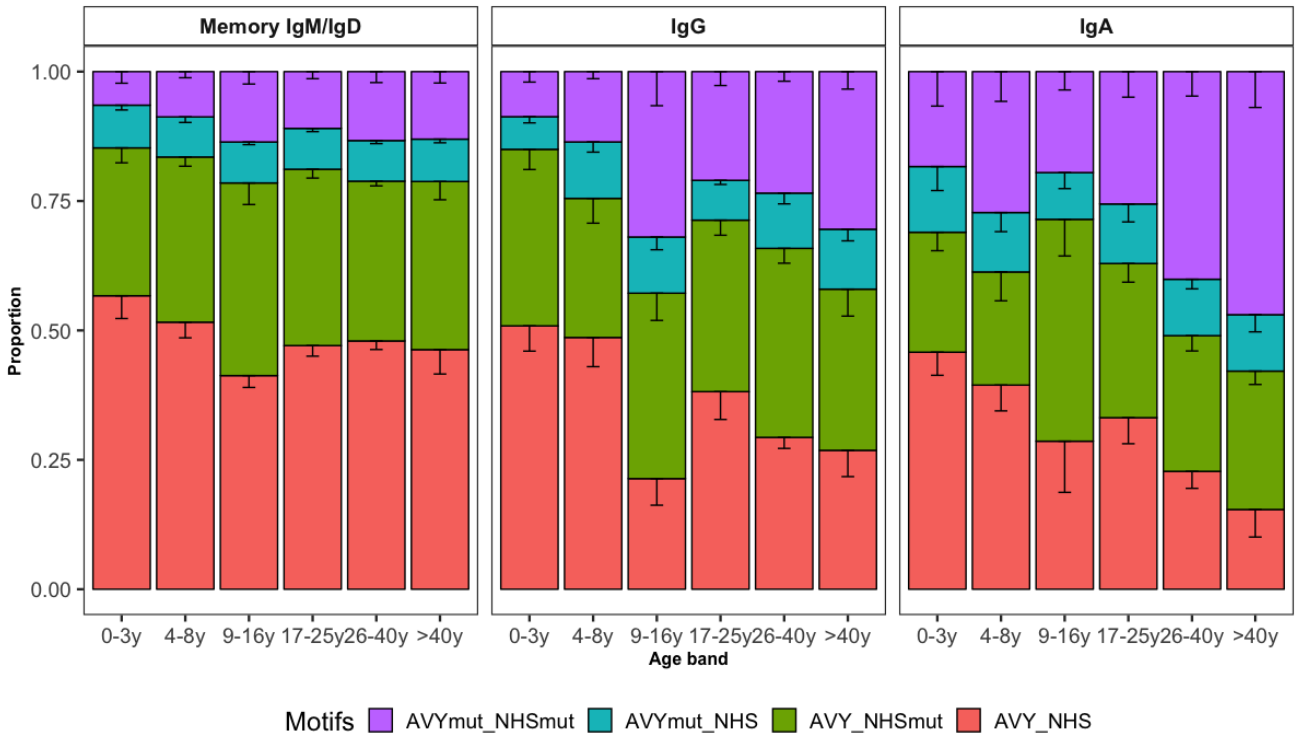


Figure 7: VH4-34 motifs. Bar plots represent the proportion of sequences with mutated AVY and/or NHS motifs in IgD, IgM, IgG and IgA. Error bars indicate SEM. Proportion of sequences with both unmutated motifs decreases with age.

Discussion

In this study, we found an extensive maturation of the B-cell system in the first 10 years of life consistent with what would be expected with cumulative antigen exposure. Further antibody repertoire alterations continue to be made thereafter, although at a lower rate. The results presented here constitute by far the most in-depth evaluation of the maturation of the B-cell system with age. This study also provides a detailed reference data set of isotype subclass-specific BCR repertoires of healthy individuals across a relevant age range and stresses the importance of using well-selected, age-appropriate controls in future studies.

Previous studies have suggested that immunoglobulin gene usage is strongly genetically determined as it was conserved between monozygotic twins and across multiple time points within a given individual (7, 33). We found age-dependent alterations in both V family and J gene usage in antigen-experienced repertoires suggesting polyclonal negative selection of V1- and J6-containing B cells during maturation of the adaptive immune system. However, here we also saw that V family gene usage changed in naïve repertoires that are supposedly unaffected by antigen exposure and not subject to selection pressure, indicating preferential development of V1-bearing B cells in young children. Although the potential benefit and mechanism behind these age-related V family gene alterations remain unclear, these findings suggest that immunoglobulin gene usage in developing B cells is less conserved than previously thought.

In line with earlier findings (11, 45, 46), we observed extensive maturation of antigen-experienced repertoires characterized by higher levels of somatic hypermutation and strong positive selection in older individuals. Of note, detailed analysis made it possible to investigate characteristics of mutated IgM/D transcripts separately, which were observed at a higher frequency and with a greater number of mutations in older individuals. These findings indicate that the pool of circulating peripheral blood naïve B cells is continuously diminishing with age, possibly contributing to a decreasing capacity to effectively respond to novel antigens in older individuals (47). We also observed a substantially higher proportion of unmutated IgA/G transcripts in young children compared with adults (48), which has not yet been recognized. These results confirm previous *in vitro* studies (49) demonstrating that class-switch recombination and somatic hypermutation can occur independently and confirm class-switching to be an important element of B-cell responses in young children.

Along with other characteristics indicative of antigen-driven maturation we found that the proportion of sequences with structures differing from germline greatly increased with age. To date, there is limited information on modelled antibody structures derived from high-throughput adaptive immune receptor sequencing data (50, 51). In line with measures of antigen-driven selection, there was a positive linear relationship between number of mutations and structural alterations of antigen-experienced sequences indicating that alteration of the three-dimensional structure is important to achieve high specificity and affinity of the antibody. By annotating individual sequences with PDB codes, we were able to investigate commonalities of CDR3 structures between individuals and assess age-dependent changes in PDB (structure) frequencies. Future work, such as the investigation of PDB usage in patients with immune disorders, will help determine how structures can be used to assess global immune competence.

We found an increase in the usage of IgA2/IgG2 transcripts with age, similar to what has been seen in a recent study on the isotype subclasses surface expression of peripheral blood B cells (52). Although human IgG subclasses have been extensively studied (53), there is limited information on the functional difference between the two IgA subclasses, whose structures mainly differ in the length of the hinge region (54). IgG2 has been implicated in the immune responses to capsular polysaccharides of bacteria such as *S. pneumoniae* that are commonly colonizing the oropharynx of young children and thereby induce polysaccharide-specific serum antibody (55). Our findings also match the sequential model proposed for CSR: with age, and after multiple encounter with the same antigen, class-switched memory B cells re-enter the

germinal center to undergo a second round of CSR and switch towards more downstream constant region genes (56).

The majority of early immature human B cells display self-reactivity and although most of these are removed during B-cell development, a substantial proportion of mature B cells may still be directed against autoantigens (39). Antibodies encoded by germline VH4-34 are intrinsically self-reactive antibodies mediated by a hydrophobic patch and a glycosylation sequon (42, 44). Unmutated VH4-34 transcripts are more common in naïve than antigen-experienced repertoires as receptor editing of these antibodies drives specificity away from self (43, 57). In contrast to adults, we found that a substantial proportion of VH4-34 IgG and IgA transcripts from children are unmutated, with frequencies gradually decreasing with age. Previous work has shown that germline VH4-34-expressing IgG B cells recognized antigens from commensal gut bacterial (57) and hence, the higher frequency of these cells in children may be related to ongoing immune responses against gut pathogens in this age group.

This study used Ig-seq technology coupled with bioinformatic methods to study in detail the BCR repertoires of healthy individuals and investigate the effect of age on repertoire characteristics. We chose a cross-sectional study design and – although unlikely – can therefore not exclude that longitudinal assessment of maturation in individuals may differ from the presented findings. For practical reasons, the number of input cells was variable between study participants, which resulted in variable sequence numbers per sample. Although subsampling is possible, this would result in removal of large numbers of sequences thereby reducing depth and increasing variability of outcome measures.

We were able to map in detail the characteristics, magnitude and speed of age-dependent maturation of BCR repertoires. This now allows comparisons to be made in the BCR repertoires of healthy individuals to individuals with altered immune states such as primary immunodeficiency (4) or infectious disease (58, 59). By elucidating patterns that are associated with cumulative antigen exposure and an evolving immune system, this research offers novel insights into the interaction between antigen and the responding adaptive immune system. The mechanisms behind the development of clinically relevant autoimmunity is still poorly understood and the findings in this study show a substantial intrinsic capacity to produce self-reactive B cells, which may be essential to achieve the diversity needed for the defense against commensal pathogens in early life.

In summary, by studying the maturation of the healthy BCR repertoire with age, we found characteristics indicative of a maturing B-cell system consisting of alterations in immunoglobulin gene usage, increased levels of SHM associated with strong positive selection, and canonical class usage that differed considerably from germline structures. Repertoires from older individuals more frequently contained transcripts using more downstream constant region genes that are involved in the immune response to polysaccharide antigens. With accumulating mutations, germline-encoded self-reactive transcripts were seen less with advancing age indicating a possible role of self-reactive B-cells in the developing immune system. This study also provides a reference data set of isotype subclass-specific BCR repertoires and stresses the importance of using well-selected, age-appropriate controls in future studies.

Author contributions

JT designed and supervised the study, oversaw analyses, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. The first draft was written by JT and MG. VvN, JDG and MG processed samples and prepared sequencing libraries. MG, JDG, AK and JT performed bioinformatic analysis, revised the manuscript and approved the final version. JPS, EM, DFK and CMD contributed to manuscript revision, and approved the final version.

Competing interests

None of the authors have declared any conflict of interest.

References

1. Reich, N. C. 2008. Janeway's Immunobiology . Seventh Edition. By *Kenneth Murphy, Paul Travers,* and *, Mark Walport; contributions by, Michael Ehrenstein, Claudia Mauri, Allan Mowat,* and *, Andrey Shaw.* Garland Science *. New York: Taylor &amp*, 7th ed. (K. Murphy, P. Travers, and M. Walport, eds). Garland Science, New York and London.

2. Rawlings, D. J., G. Metzler, M. Wray-Dutra, and S. W. Jackson. 2017. Altered B cell signalling in autoimmunity. *Nat. Rev. Immunol.* 17: 421–436.

3. Hoffman, W., F. G. Lakkis, and G. Chalasani. 2016. B Cells, Antibodies, and More. *Clin. J. Am. Soc. Nephrol.* 11: 137–54.

4. Ghraichy, M., J. D. Galson, D. F. Kelly, and J. Trück. 2018. B-cell receptor repertoire sequencing in patients with primary immunodeficiency: a review. *Immunology* 153: 145–160.

5. Bashford-Rogers, R. J. M., K. G. C. Smith, and D. C. Thomas. 2018. Antibody repertoire analysis in polygenic autoimmune diseases. *Immunology* .

6. Jiang, N., J. He, J. A. Weinstein, L. Penland, S. Sasaki, X. S. He, C. L. Dekker, N. Y. Zheng, M. Huang, M. Sullivan, P. C. Wilson, H. B. Greenberg, M. M. Davis, D. S. Fisher, and S. R. Quake. 2013. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* 5.

7. Galson, J. D., J. Trück, E. A. Clutterbuck, A. Fowler, V. Cerundolo, A. J. Pollard, G. Lunter, and D. F. Kelly. 2016. B-cell repertoire dynamics after sequential hepatitis B vaccination and evidence for cross-reactive B-cell activation. *Genome Med.* 8: 68.

8. Galson, J. D., J. Trück, A. Fowler, M. Münz, V. Cerundolo, A. J. Pollard, G. Lunter, and D. F. Kelly. 2015. In-Depth Assessment of Within-Individual and Inter-Individual Variation in the B Cell Receptor Repertoire. *Front. Immunol.* 6: 1–13.

9. Galson, J. D., E. a Clutterbuck, J. Trück, M. N. Ramasamy, M. Münz, A. Fowler, V. Cerundolo, A. J. Pollard, G. Lunter, and D. F. Kelly. 2015. BCR repertoire sequencing: different patterns of B-cell activation after two Meningococcal vaccines. *Immunol. Cell Biol.* 93: 885–895.

10. Kovaltsuk, A., J. Leem, S. Kelm, J. Snowden, C. M. Deane, and K. Krawczyk. 2018. Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *J. Immunol.* .

11. IJspeert, H., P. A. van Schouwenburg, D. van Zessen, I. Pico-Knijnenburg, G. J. Driessen, A. P. Stubbs, and M. van der Burg. 2016. Evaluation of the antigen-experienced B-cell receptor repertoire in healthy children and adults. *Front. Immunol.* 7.

12. Comans-Bitter, W. M., R. De Groot, R. Van den Beemd, H. J. Neijens, W. C. J. Hop, K. Groeneveld, H. Hooijkaas, and J. J. M. Van Dongen. 1997. Immunophenotyping of blood lymphocytes in childhood: Reference values for lymphocyte subpopulations. *J. Pediatr.* 130: 388–393.

13. Vander Heiden, J. A., G. Yaari, M. Uduman, J. N. H. Stern, K. C. O'Connor, D. A. Hafler, F. Vigneault, and S. H. Kleinstein. 2014. PRESTO: A toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* 30: 1930–1932.

14. Gupta, N. T., J. A. Vander Heiden, M. Uduman, D. Gadala-Maria, G. Yaari, and S. H. Kleinstein. 2015. Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* 31: 3356–3358.

15. Ye, J., N. Ma, T. L. Madden, and J. M. Ostell. 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* 41.

16. Lunter, G., and M. Goodson. 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* 21: 936–939.

17. Yaari, G., M. Uduman, and S. H. Kleinstein. 2012. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Res.* .

18. Stern, J. N. H., G. Yaari, J. A. Vander Heiden, G. Church, W. F. Donahue, R. Q. Hintzen, A. J. Huttner, J. D. Laman, R. M. Nagra, A. Nylander, D. Pitt, S. Ramanan, B. A. Siddiqui, F. Vigneault, S. H. Kleinstein, D. A. Hafler, and K. C. O'Connor. 2014. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci. Transl. Med.* .

19. Dunbar, J., and C. M. Deane. 2015. ANARCI: Antigen receptor numbering and receptor classification. *Bioinformatics* .

20. Wong, W. K., G. Georges, F. Ros, S. Kelm, A. P. Lewis, B. Taddese, J. Leem, and C. M. Deane. 2018. SCALOP: sequence-based antibody canonical loop structure annotation. *Bioinformatics* .

21. Choi, Y., and C. M. Deane. 2010. FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins Struct. Funct. Bioinforma.* .

22. Deane, C. M. 2001. CODA: A combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci.* .

23. Kovaltsuk, A., K. Krawczyk, S. Kelm, J. Snowden, and C. M. Deane. 2018. Filtering Next-Generation Sequencing of the Ig Gene Repertoire Data Using Antibody Structural Information. *J. Immunol.* .

24. Lefranc, M. P., C. Pommié, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet, and G. Lefranc. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.* .

25. Nowak, J., T. Baker, G. Georges, S. Kelm, S. Klostermann, J. Shi, S. Sridharan, and C. M. Deane. 2016. Length-independent structural similarities enrich the antibody CDR canonical class model. *MAbs* 8: 751–760.

26. Dunbar, J., K. Krawczyk, J. Leem, T. Baker, A. Fuchs, G. Georges, J. Shi, and C. M. Deane. 2014. SAbDab: The structural antibody database. *Nucleic Acids Res.* .

27. Chothia, C., and A. M. Lesk. 1987. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* .

28. Wang, D., L. H. Lai, P. Yan, and P. S. Dong. 2017. Correlation between blood asymmetric dimethylarginine level and the complications of patients with cardiovascular diseases. *J. Biol. Regul. Homeost. Agents* 31: 133–139.

29. Ginestet, C. 2011. ggplot2: Elegant Graphics for Data Analysis. *J. R. Stat. Soc. Ser. A (Statistics Soc.* .

30. Kassambara, A. 2018. ggpubr: “ggplot2” Based Publication Ready Plots. *R Packag. version 0.1.8.* .

31. Luo, S., J. A. Yu, and Y. S. Song. 2016. Estimating Copy Number and Allelic Variation at the Immunoglobulin Heavy Chain Locus Using Short Reads. *PLoS Comput. Biol.* 12.

32. Rubelt, F., C. E. Busse, S. A. C. Bukhari, J.-P. Bürckert, E. Mariotti-Ferrandiz, L. G. Cowell, C. T. Watson, N. Marthandan, W. J. Faison, U. Hershberg, U. Laserson, B. D. Corrie, M. M. Davis, B. Peters, M.-P. Lefranc, J. K. Scott, F. Breden, E. T. Luning Prak, and S. H. Kleinstein. 2017. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nat. Immunol.* .

33. Glanville, J., T. C. Kuo, H.-C. von Budingen, L. Guey, J. Berka, P. D. Sundar, G. Huerta, G. R. Mehta, J. R. Oksenberg, S. L. Hauser, D. R. Cox, A. Rajpal, and J. Pons. 2011. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl. Acad. Sci.* 108: 20066–20071.

34. IJspeert, H., J. Rozmus, K. Schwarz, R. L. Warren, D. Van Zessen, R. A. Holt, I. Pico-Knijnenburg, E. Simons, I. Jerchel, A. Wawer, M. Lorenz, T. Patiro□lu, H. H. Akar, R. Leite, N. S. Verkaik, A. P. Stubbs, D. C. Van Gent, J. J. M. Van Dongen, and M. Der Van Burg. 2016. XLF deficiency results in reduced N-nucleotide addition during V(D)J recombination. *Blood* 128: 650–659.

35. Donisi, P. M., N. Di Lorenzo, M. Riccardi, A. Paparella, C. Sarpellon, S. Zupo, G. Bertoldero, C. Minotto, and V. Stracca-Pansa. 2006. Pattern and distribution of immunoglobulin VH gene usage in a cohort of B-CLL patients from a northeastern region of Italy. *Diagnostic Mol. Pathol.* .

36. Widhopf, G. F., L. Z. Rassenti, T. L. Toy, J. G. Gribben, W. G. Wierda, and T. J. Kipps. 2004. Chronic lymphocytic leukemia B cells of more than 1% of patients express virtually identical immunoglobulins. *Blood* .

37. North, B., A. Lehmann, and R. L. Dunbrack. 2011. A new clustering of antibody CDR loop conformations. *J. Mol. Biol.* 406: 228–256.

38. Tsioris, K., N. T. Gupta, A. O. Ogunniyi, R. M. Zimnisky, F. Qian, Y. Yao, X. Wang, J. N. H. Stern, R. Chari, A. W. Briggs, C. R. Clouser, F. Vigneault, G. M. Church, M. N. Garcia, K. O. Murray, R. R. Montgomery, S. H. Kleinstein, and J. C. Love. 2015. Neutralizing antibodies against West Nile virus identified directly from human B cells by single-cell analysis and next generation sequencing. *Integr. Biol. (United Kingdom)* .

39. Wardemann, H., S. Yurasov, A. Schaefer, J. W. Young, E. Meffre, and M. C. Nussenzweig. 2003. Predominant autoantibody production by early human B cell precursors. *Science* 301: 1374–1377.

40. Larimore, K., M. W. McCormick, H. S. Robins, and P. D. Greenberg. 2012. Shaping of human germline IgH repertoires revealed by deep sequencing. *J. Immunol.* 189: 3221–30.

41. Pugh-Bernard, A. E., G. J. Silverman, A. J. Cappione, M. E. Villano, D. H. Ryan, R. A. Insel, and I. Sanz. 2001. Regulation of inherently autoreactive VH4-34 B cells in the maintenance of human B cell tolerance. *J. Clin. Invest.* 108: 1061–1070.

42. Potter, K. N., P. Hobby, S. Klijn, F. K. Stevenson, and B. J. Sutton. 2002. Evidence for involvement of a hydrophobic patch in framework region 1 of human V4-34-encoded Igs in recognition of the red blood cell I antigen. *J. Immunol.* 169: 3777–3782.

43. Reed, J. H., J. Jackson, D. Christ, and C. C. Goodnow. 2016. Clonal redemption of autoantibodies by somatic hypermutation away from self-reactivity during human immunization. *J. Exp. Med.* 213: 1255–1265.

44. Sabouri, Z., P. Schofield, K. Horikawa, E. Spierings, D. Kipling, K. L. Randall, D. Langley, B. Roome, R. Vazquez-Lombardi, R. Rouet, J. Hermes, T. D. Chan, R. Brink, D. K. Dunn-Walters, D. Christ, and C. C. Goodnow. 2014. Redemption of autoantibodies on anergic B cells by variable-region glycosylation and mutation away from self-reactivity. *Proc. Natl. Acad. Sci.* 111: E2567–E2575.

45. Tabibian-Keissar, H., L. Hazanov, G. Schiby, N. Rosenthal, A. Rakovsky, M. Michaeli, G. L. Shahaf, Y. Pickman, K. Rosenblatt, D. Melamed, D. Dunn-Walters, R. Mehr, and I. Barshack. 2016. Aging affects B-cell antigen receptor repertoire diversity in primary and secondary lymphoid tissues. *Eur. J. Immunol.* .

46. Schatorjé, E. J. H., G. J. Driessen, R. W. N. M. van Hout, M. van der Burg, and E. de Vries. 2014. Levels of somatic hypermutations in B cell receptors increase during childhood. *Clin. Exp. Immunol.* .

47. Siegrist, C. A., and R. Aspinall. 2009. B-cell responses to vaccination at the extremes of age. *Nat. Rev. Immunol.* 9: 185–194.

48. Fecteau, J. F., G. Cote, and S. Neron. 2006. A New Memory CD27-IgG+ B Cell Population in Peripheral Blood Expressing VH Genes with Low Frequency of Somatic Mutation. *J. Immunol.* 177: 3728–3736.

49. Nagumo, H., K. Agematsu, N. Kobayashi, K. Shinozaki, S. Hokibara, H. Nagase, M. Takamoto, K. Yasui, K. Sugane, and A. Komiyama. 2002. The different process of class switching and somatic hypermutation; a novel analysis by CD27-naive B cells. *Blood* 99: 567–575.

50. Kovaltsuk, A., K. Krawczyk, J. D. Galson, D. F. Kelly, C. M. Deane, and J. Trück. 2017. How B-cell receptor repertoire sequencing can be enriched with structural antibody data. *Front. Immunol.* 8: 1753.

51. Krawczyk, K., S. Kelm, A. Kovaltsuk, J. D. Galson, D. Kelly, J. Trück, C. Regep, J. Leem, W. K. Wong, J. Nowak, J. Snowden, M. Wright, L. Starkie, A. Scott-Tucker, J. Shi, and C. M. Deane. 2018. Structurally mapping antibody repertoires. *Front. Immunol.* .

52. Blanco, E., M. Pérez-Andrés, S. Arriba-Méndez, T. Contreras-Sanfeliciano, I. Criado, O. Pelak, A. Serra-Caetano, A. Romero, N. Puig, A. Remesal, J. Torres Canizales, E. López-Granados, T. Kalina, A. E. Sousa, M. van Zelm, M. van der Burg, J. J. M. van Dongen, and A. Orfao. 2018. Age-associated distribution of normal B-cell and plasma cell subsets in peripheral blood. *J. Allergy Clin. Immunol.* 141: 2208–2219.e16.

53. Vidarsson, G., G. Dekkers, and T. Rispens. 2014. IgG subclasses and allotypes: From structure to effector functions. *Front. Immunol.* 5: 520.

54. Woof, J. M., and M. A. Kerr. 2004. IgA function - Variations on a theme. *Immunology* 113: 175–177.

55. Turner, P., C. Turner, N. Green, L. Ashton, E. Lwe, A. Jankhot, N. P. Day, N. J. White, F. Nosten, and D. Goldblatt. 2013. Serum antibody responses to pneumococcal colonization in the first 2 years of life: Results from an SE Asian longitudinal cohort study. *Clin. Microbiol. Infect.* 19: 1–8.

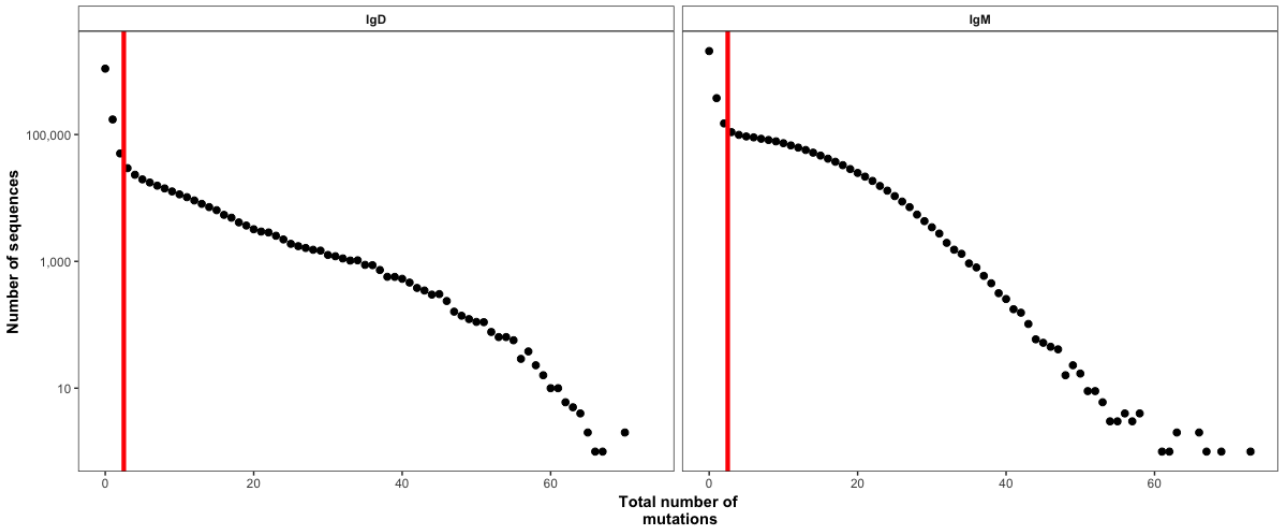
56. Xiong, H., J. Dolpady, M. Wabl, M. A. Curotto de Lafaille, and J. J. Lafaille. 2012. Sequential class switching is required for the generation of high affinity IgE antibodies. *J. Exp. Med.* .

57. Schickel, J.-N., S. Glauzy, Y.-S. Ng, N. Chamberlain, C. Massad, I. Isnardi, N. Katz, G. Uzel, S. M. Holland, C. Picard, A. Puel, J.-L. Casanova, and E. Meffre. 2017. Self-reactive VH4-34–expressing IgG B cells recognize commensal bacteria. *J. Exp. Med.* 214: 1991–2003.

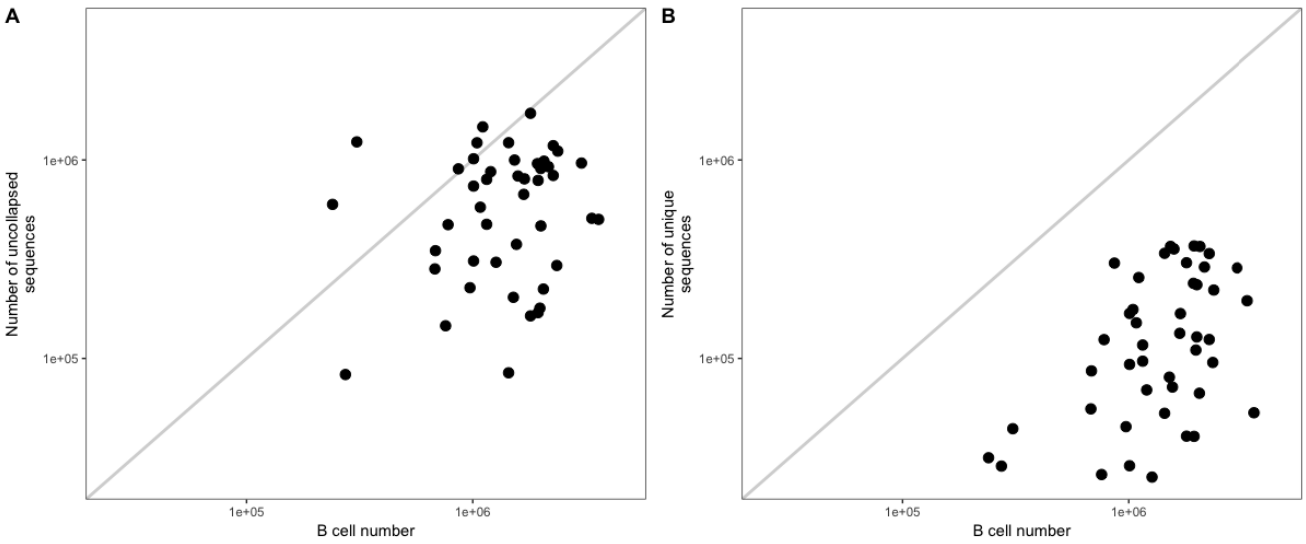
58. Hou, D., C. Chen, E. J. Seely, S. Chen, and Y. Song. 2016. High-throughput sequencing-based immune repertoire study during infectious disease. *Front. Immunol.* 7: 336.

59. Burkholder, W. F., E. W. Newell, M. Poidinger, S. Chen, and K. Fink. 2017. Deep Sequencing in Infectious Diseases: Immune and Pathogen Repertoires for the Improvement of Patient Outcomes. *Front. Immunol.* 8: 593.

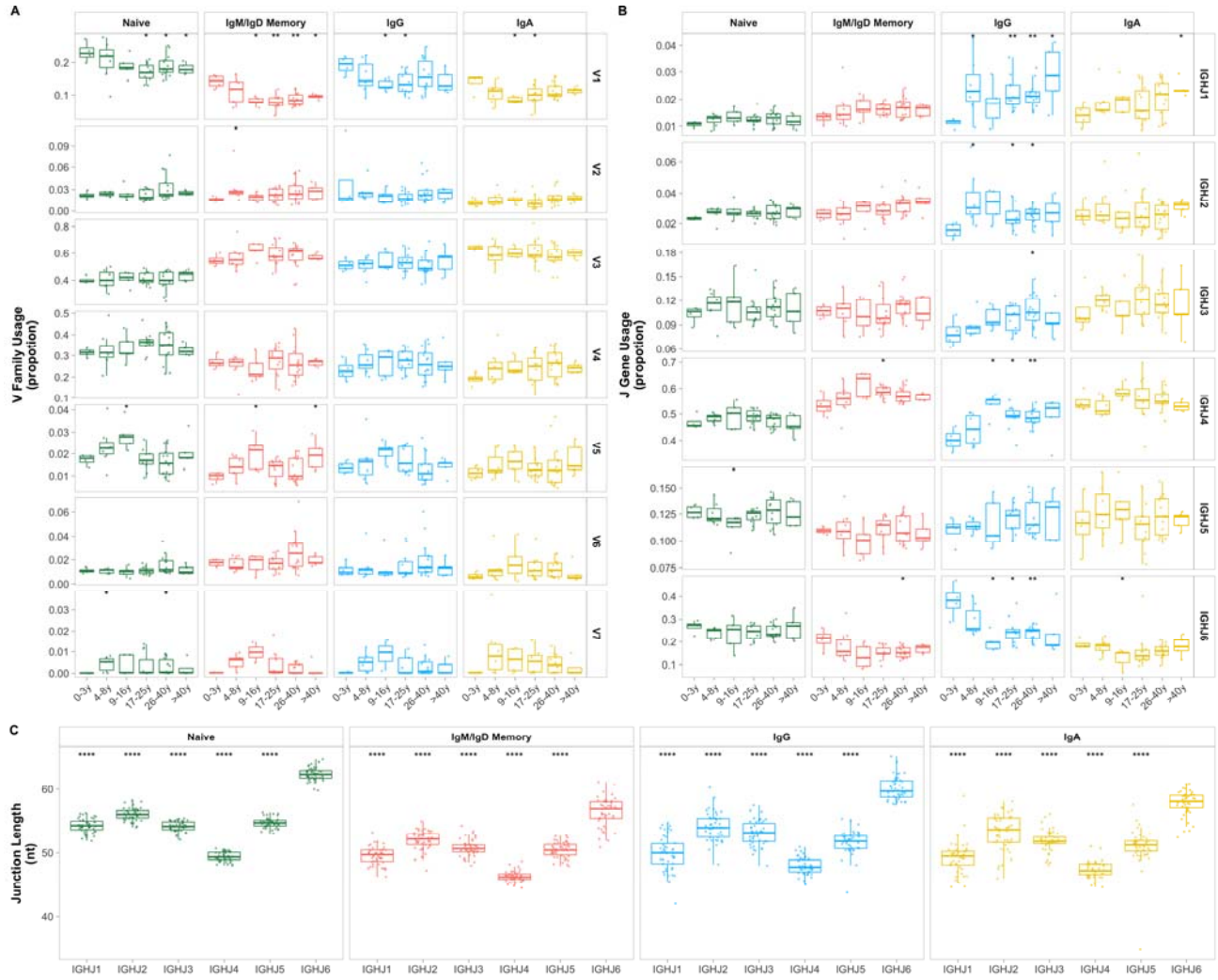
Supplementary information



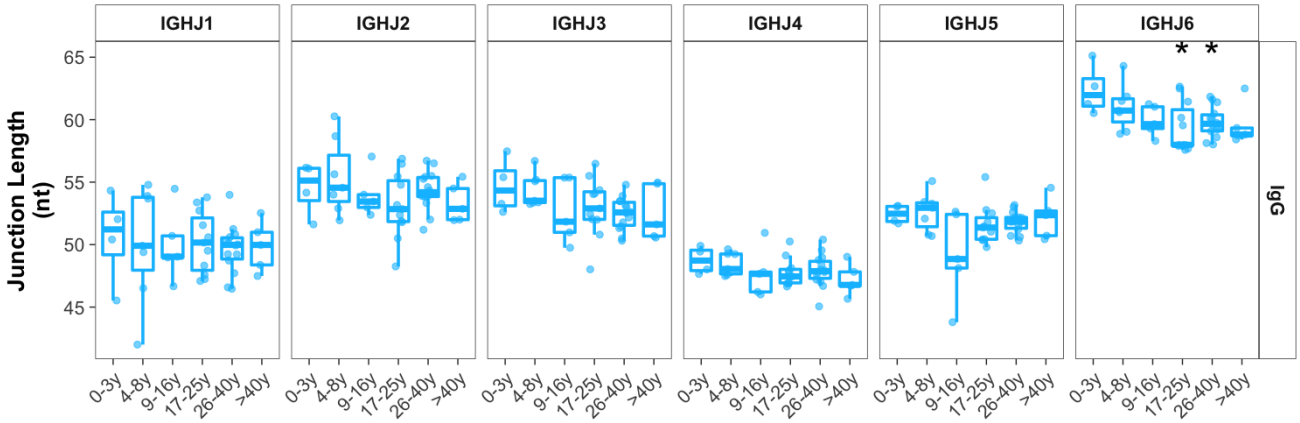
Supplementary figure 1 Distribution of somatic hypermutation in all IgD and IgM transcripts. The vertical line indicates the threshold chosen (mutation n between 2 and 3) to separate naïve and memory repertoires for IgM and IgD sequences.



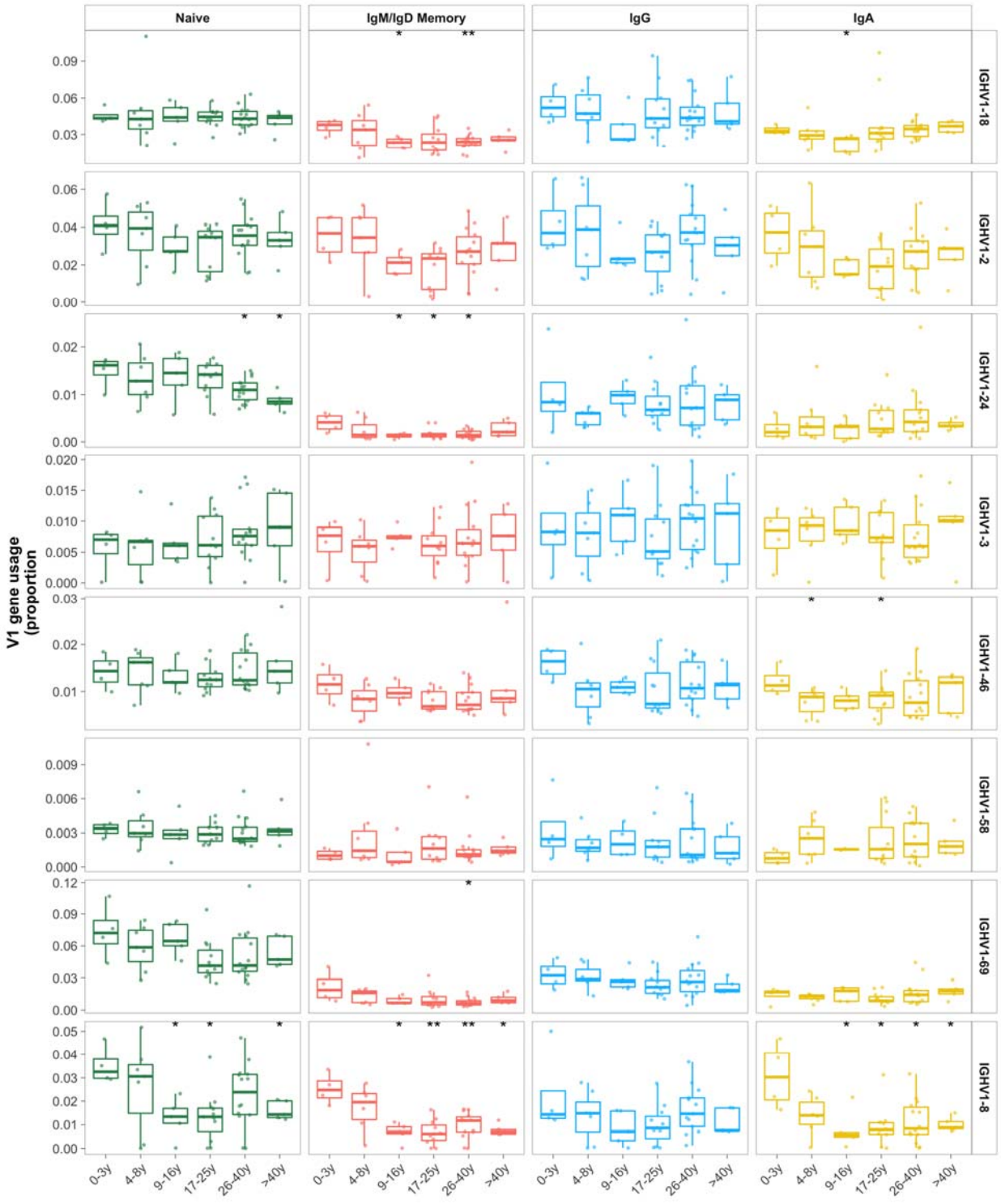
Supplementary figure 2: Correlation between cell number in a sample, and the number of sequences for that sample *A* before and *B* after collapsing. The B-cell number was either based on actual counts or estimated using PBMC counts and the median percentage of age-dependent reference values.



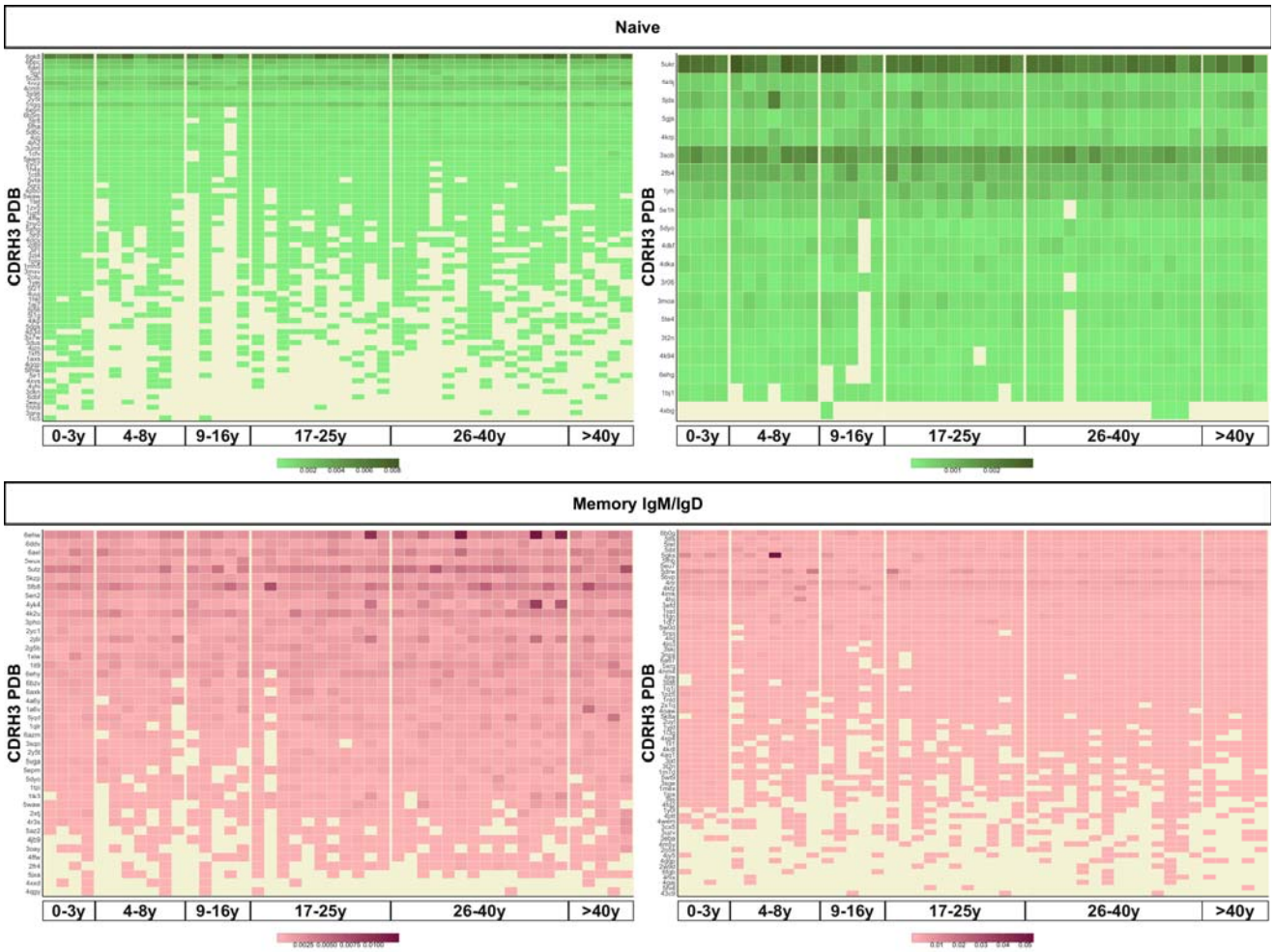
Supplementary figure 3 *A* V family and *B* J gene usage by age band. Comparison of each age group to the 0-3y group was performed using the Wilcoxon test. *C* IGHJ6 transcripts show significantly longer junctions. Comparison of each gene to IGHJ6 was performed using the Wilcoxon test. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001



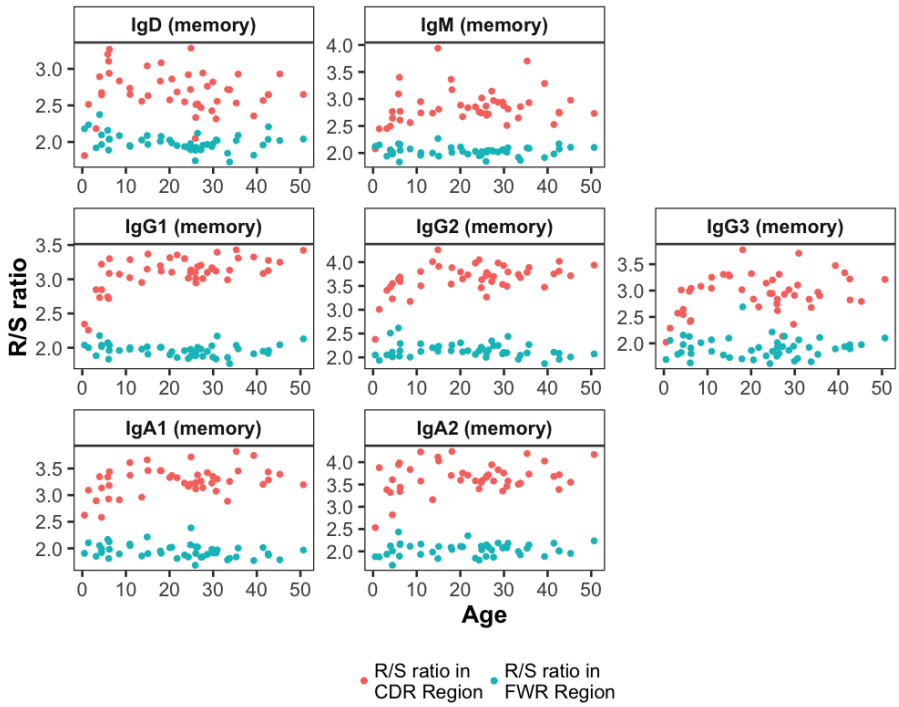
Supplementary Figure 4: Junction length decrease in IgG transcripts is still apparent when normalizing for J gene and is only significant in transcripts with IGHJ6. Comparison of each age group to the 0-3y group was performed using the Wilcoxon test. *p<0.05

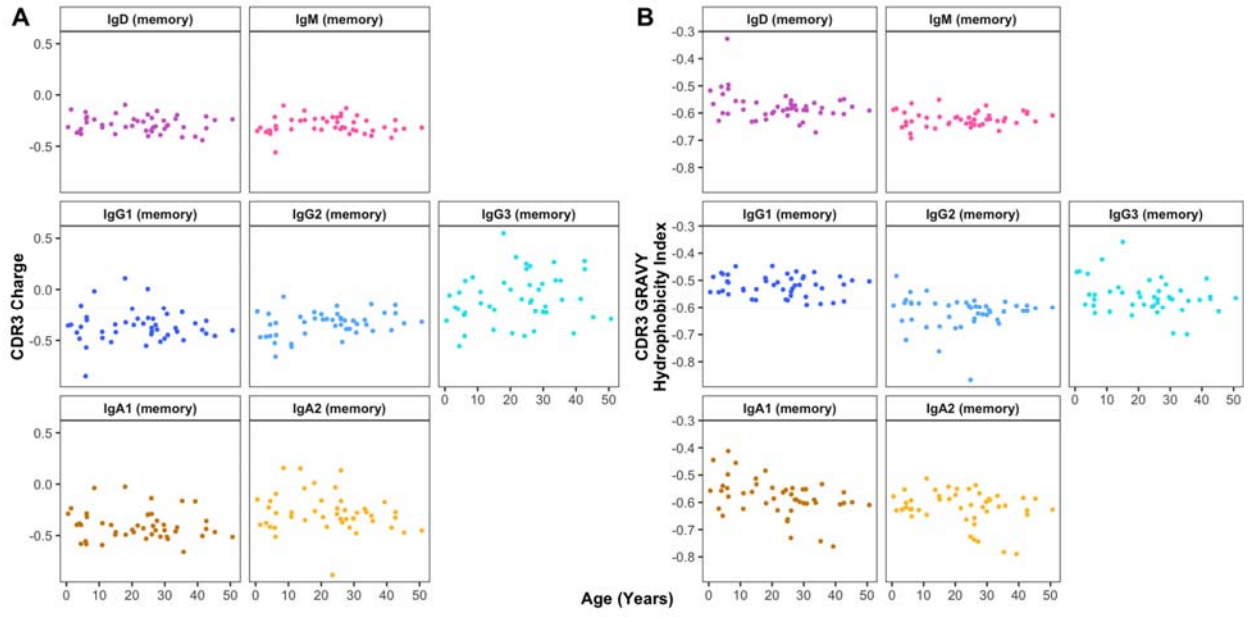


Supplementary figure 5 Proportion of the top 8 V1 family genes by age band. The decrease seen in V1 family usage is a result of a decrease in multiple individual genes. Comparison of each age group to the 0-3y group was performed using the Wilcoxon test. *p<0.05, **p<0.01



Supplementary Figure 6 The structural composition of the naïve and IgM/IgD memory repertoires. PDB codes that have a positive correlation (left) and a negative correlation (right) with age and with a correlation p-value <0.05 are shown. Out of 2040 unique naïve PDBs, 68 (0.03%) and 20 (0.01%) were positively and negatively correlated with age, respectively. Out of 1990 IgM/D memory PDBs, 42 (0.02%) and 66 (0.03%) were positively and negatively correlated with age, respectively. Samples are ordered by age and PDB codes are ordered by sharedness across individuals.





Supplementary figure 7 R/S ratio is lower in FWR region compared to CDR region and does not correlate with age,

Supplementary figure 8 *A* CDR3 charge and *B* CDR3 hydrophobicity index do not correlate with age in healthy controls.